Title

Visualizing 'omic feature rankings and log-ratios using Qurro.

Journal

NAR genomics and bioinformatics, 2(2)

ISSN

2631-9268

Authors

Fedarko, Marcus W
Martino, Cameron
Morton, James T
et al.

Publication Date

2020-06-01

DOI

10.1093/nargab/lqaa023

Peer reviewed

Published online 28 April 2020

NAR Genomics and Bioinformatics, 2020, Vol. 2, No. 2 **1**
doi: 10.1093/nargab/lqaa023

Visualizing 'omic feature rankings and log-ratios using Qurro

Marcus W. Fedarko [1,2], Cameron Martino[2,3], James T. Morton[4], Antonio González[5], Gibraan Rahman[3], Clarisse A. Marotz[6], Jeremiah J. Minich[7], Eric E. Allen[2,7,8] and Rob Knight [1,2,5,9,*]

[1]Department of Computer Science and Engineering, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA, [2]Center for Microbiome Innovation, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA, [3]Bioinformatics and Systems Biology Program, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA, [4]Flatiron Institute, Simons Foundation, 162 Fifth Avenue, New York City, NY 10010, USA, [5]Department of Pediatrics, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA, [6]Department of Biomedical Sciences, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA, [7]Marine Biology Research Division, Scripps Institution of Oceanography, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA, [8]Department of Biological Sciences, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA and [9]Department of Bioengineering, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

Received November 29, 2019; Revised March 11, 2020; Editorial Decision March 22, 2020; Accepted March 31, 2020

ABSTRACT

Many tools for dealing with compositional ''omics' data produce feature-wise values that can be ranked in order to describe features' associations with some sort of variation. These values include differentials (which describe features' associations with specified covariates) and feature loadings (which describe features' associations with variation along a given axis in a biplot). Although prior work has discussed the use of these 'rankings' as a starting point for exploring the log-ratios of particularly high- or low-ranked features, such exploratory analyses have previously been done using custom code to visualize feature rankings and the log-ratios of interest. This approach is laborious, prone to errors and raises questions about reproducibility. To address these problems we introduce Qurro, a tool that interactively visualizes a plot of feature rankings (a 'rank plot') alongside a plot of selected features' log-ratios within samples (a 'sample plot'). Qurro's interface includes various controls that allow users to select features from along the rank plot to compute a log-ratio; this action updates both the rank plot (through highlighting selected features) and the sample plot (through displaying the current log-ratios of samples). Here, we demonstrate how this unique interface helps users explore feature rankings and log-ratios simply and effectively.

INTRODUCTION

High-throughput sequencing and metabolomics data detailing the organisms, genes or molecules identified within a microbial sample are inherently compositional (1,2): that is, absolute abundances are often inaccessible and only relative information can be obtained from the data. These data must be interpreted accordingly. Performing a differential abundance analysis in a dataset generally requires selecting a 'reference frame' (denominator) for log-ratio analysis, then relating the resulting log-ratios to sample metadata (1,2). Critically, how to best select such a 'reference frame' is an open question. The implicit use of different references across different studies can be a cause of irreproducible findings (2).

Various tools for differential abundance analyses including but not limited to ALDEx2 (3) and Songbird (2) can produce *differentials*, which describe the (estimated) log-fold change in relative abundance for features in a dataset with respect to certain covariate(s) (2). Similarly, tools like DEICODE (4) can produce *feature loadings* that characterize features' impacts in a compositional biplot (5). Differentials and feature loadings alike can be sorted numerically and used as *feature rankings*, and this representation provides relative information about features' associations with some sort of variation in a dataset (2,4). The natural next

*To whom correspondence should be addressed. Tel: +1 858 822 2379; Fax: +1 858 246 1981; Email: robknight@ucsd.edu

step is to use these rankings as a guide for log-ratio analyses (e.g. by examining the log-ratios of high- to low-ranked features). However, modern studies commonly describe hundreds or thousands of observed features: manually exploring feature rankings, whether as a tabular representation or as visualized using one-off scripts, is inconvenient.

Here we present Qurro (pronounced 'churro'), a visualization tool that supports the analysis of feature log-ratios in the context of feature rankings and sample metadata. Qurro uses a two-plot interface: a 'rank plot' shows how features are differentially ranked for a selected differential or feature loading (as shown in Figures 1A and 2A), and a 'sample plot' shows log-ratios of the selected features across samples relative to selected sample metadata field(s) (as shown in Figures 1B and C, 2B and C). These plots are linked (6): selecting features for a log-ratio highlights these features in the rank plot and updates the y-axis values of samples (corresponding to the value of the currently selected log-ratio for each sample) in the sample plot. This interface is intended to make it easy for researchers to explore log-ratios in a dataset, using feature rankings as a starting point.

Due to its unique display, and the availability of multiple controls for feature selection and plot customization, Qurro simplifies compositional data analyses of 'omic data.

IMPLEMENTATION

Qurro's source code is released under the BSD 3-clause license and is available at https://github.com/biocore/qurro.

Qurro's codebase includes a Python 3 program that generates a visualization and the HTML/JavaScript/CSS code that manages this visualization. Qurro can be used as a standalone program or as a QIIME 2 plugin (7).

Both plots in a Qurro visualization are embedded as Vega-Lite JSON specifications (8), which are generated by Altair (9) in Qurro's Python code. An advantage of Qurro's use of the Vega infrastructure is that both plots in a Qurro visualization can be customized to the user's liking in the Vega-Lite or Vega grammars. As an example of this customizability, the Vega-Lite specifications defining Figures 1A–C and 2A–C of this paper were edited programmatically in order to increase font sizes, change the number of ticks shown, etc. (Our Python script that makes these modifications is available online; please see the 'Data Availability' section.)

Code dependencies

In addition to Altair, Qurro's Python code directly relies on the BIOM format (10), Click (https://palletsprojects.com/p/click), NumPy (11), pandas (12) and scikit-bio (http://scikit-bio.org) libraries. Qurro's web code relies on Vega (13), Vega-Lite (8), Vega-Embed (https://github.com/vega/vega-embed), RequireJS (https://requirejs.org), jQuery (https://jquery.com), DataTables (https://datatables.net), Bootstrap (https://getbootstrap.com), Bootstrap Icons (https://icons.getbootstrap.com), and Popper.js (https://popper.js.org).

CASE STUDY: THE GILLS OF *SCOMBER JAPONICUS*

To demonstrate the utility of Qurro on a dataset with clear 'signals,' we applied it to an extant dataset of V4-region 16S rRNA sequencing data from Pacific chub mackerel (*Scomber japonicus*) and environmental samples (14). This dataset, currently described in a preprint, includes samples taken from five *S. japonicus* body sites (digesta, GI, gill, pyloric caeca and skin) from 229 fish captured across 38 time points in 2017, along with many seawater, marine sediment, positive/negative control and non-*S. japonicus* fish samples. A Jupyter Notebook (15) showing how we processed this dataset computationally is available online; see the 'Data Availability' section.

Sample processing and analysis

When these samples were initially sequenced, the KatharoSeq protocol (16) was followed. This led us to exclude samples with less than 1370 total counts from our analysis of this dataset.

Sequencing data (already processed using QIIME 1.9.1 (17) and Deblur (18) on Qiita (19)) were further processed and analyzed using QIIME 2 (7). Our use of Deblur outputs as the starting point in our analysis means that 'features' in our analysis of this dataset correspond to 'sub-operational-taxonomic-units' (sOTUs), although Qurro is interoperable with compositional datasets including arbitrary types of 'features.'

These sOTUs were assigned taxonomic classifications using q2-feature-classifier (20). Specifically, we extracted sequences from the SILVA 132 99% database (21) using the same forward (22) and reverse (23) primer sequences as were used for sample processing, trained a Naïve Bayes classifier on these extracted sequences, and then used this classifier (through q2-feature-classifier's `classify-sklearn` method (24)) to classify sOTUs in our dataset based on their sequences.

Due to upstream filtering steps taken in our analysis (a combination of filtering out non-*S. japonicus* and non-seawater samples, applying the aforementioned KatharoSeq sample exclusion criterion, Songbird's default `--min-feature-count` of each feature needing to be present in at least 10 samples, and Qurro's behavior of filtering out empty samples and features), 639 samples and 985 features were included in the Qurro visualization produced for this case study.

Computing 'body site' differentials

One basic question about this dataset we investigated using Qurro was of which features were associated with which *S. japonicus* body sites. To produce feature rankings accordingly, we used Songbird (2) to compute differentials detailing features' associations with samples from each of the five studied body sites, using the seawater samples in the dataset as a reference category for Songbird's internal construction of a design matrix representing the sample categories being analyzed (Supplementary Data, section 1).

In general, highly ranked features for a differential—the (estimated) log-fold change in relative abundance for a fea-

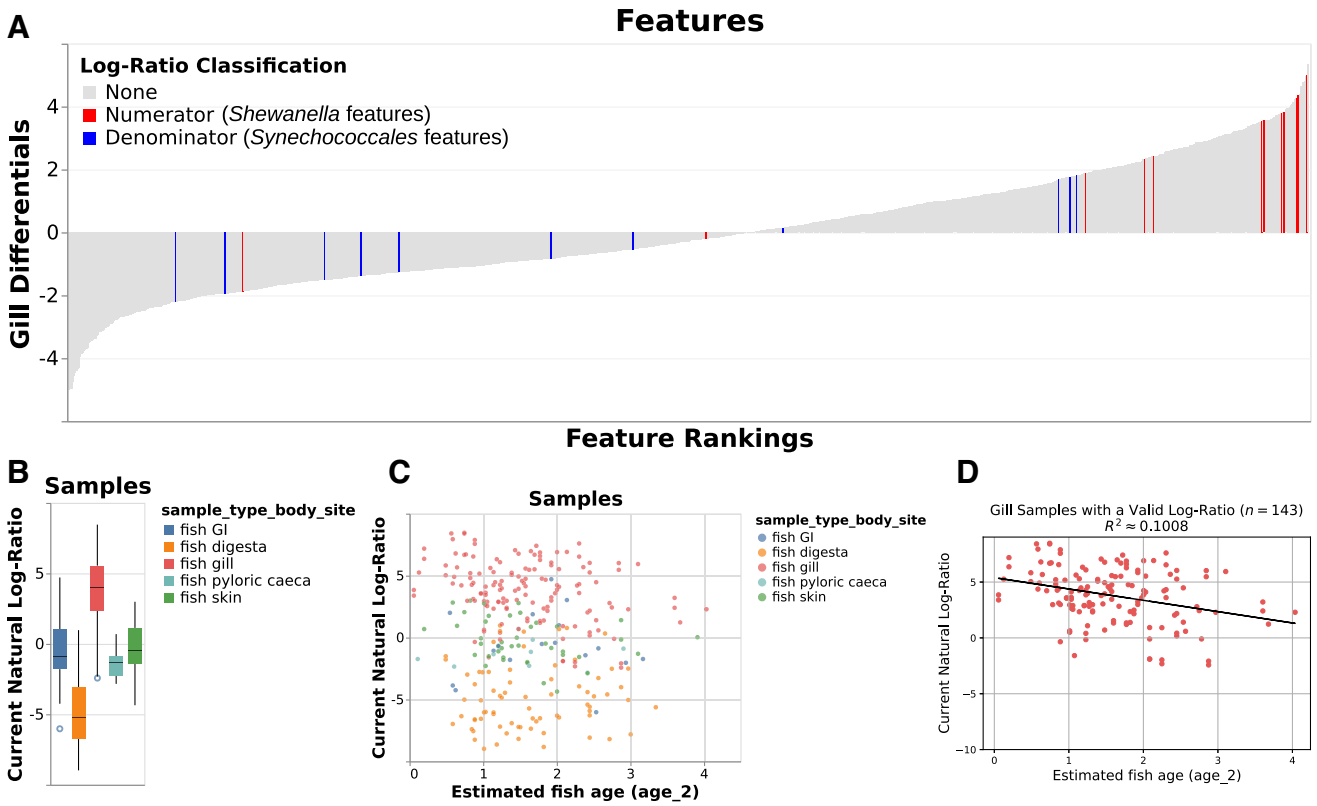

Figure 1. Various outputs from the case study showing the log-ratio of classified *Shewanella* features to classified *Synechococcales* features. (**A**) 'Rank plot' showing differentials computed based on association with gill samples, using seawater samples as a reference category in the regression. The term 'log-ratio classification' only refers to the currently selected log-ratio in the Qurro visualization: in this case, this is the log-ratio of classified *Shewanella* features to classified *Synechococcales* features. To show the rankings of these 'selected' features relative to the remaining features in the dataset, these features are colored in the rank plot: *Shewanella* features are colored in red, and *Synechococcales* features are colored in blue. The remaining features, colored gray, have a 'log-ratio classification' of `None` because they are not involved in the selected *Shewanella*-to-*Synechococcales* log-ratio. (**B**) 'Sample plot' in boxplot mode, showing samples' *Shewanella*-to-*Synechococcales* log-ratios by sample body site. Note that only 285 samples are represented in this plot; other samples were either filtered out upstream in the analysis or contained zeroes on at least one side of their log-ratio. (**C**) 'Sample plot,' showing a scatterplot of samples' selected log-ratios versus estimated fish age. Individual samples are colored by body site. As in panel B, only 285 samples are present. (**D**) Ordinary-least-squares linear regression ($R^2 \approx 0.1008$) between estimated fish age and the selected log-ratio for just the 143 gill samples shown in B and C, computed outside of Qurro using scikit-learn (24) and pandas (12) and plotted using matplotlib (30).

ture with respect to some covariate(s)—are positively associated with samples from these covariate(s), while lowly ranked features are negatively associated with these covariate(s). These differentials can therefore be thought of as a starting point for investigating differentially abundant features for particular fish body sites in this dataset.

Using Qurro to analyze differentials and log-ratios

Qurro simplifies the process of analyzing features' log-ratios in the context of these differentials. The 'rank plot' of a Qurro visualization is a bar plot where each bar corresponds to a single differentially ranked feature. The y-axis values of each feature's bar are either the estimated log-fold change values for that feature if the feature rankings are differentials (as is the case in our analysis here, and therefore in Figures 1A and 2A), or the loadings of each feature along a selected biplot axis if the feature rankings are feature loadings in a biplot. In either case, features are sorted in ascending order by these values along the rank plot's *x*-axis. The exact differential or feature loading used is configurable, so Qurro users can quickly toggle between these; for the case study Qurro visualization, this means that users can—for example—quickly switch between differentials computed based on association with skin samples to differentials computed based on association with gill samples.

Highlighting features on the rank plot. The initial study of this dataset (14) agreed with prior work (25) on the frequency of *Shewanella* spp. in the fish gill microbiome. Qurro supports searching for features using arbitrary feature metadata (e.g. taxonomic annotations), and using this functionality to highlight *Shewanella* spp. on the rank plot of gill differentials (Supplementary Data, section 2) corroborates these findings: as Figure 1A shows, the majority of identified *Shewanella* spp. are highly ranked for the gill differentials relative to the other features in this dataset.

Particularly high- or low-ranked features like *Shewanella* spp. can merit further examination via a log-ratio analysis

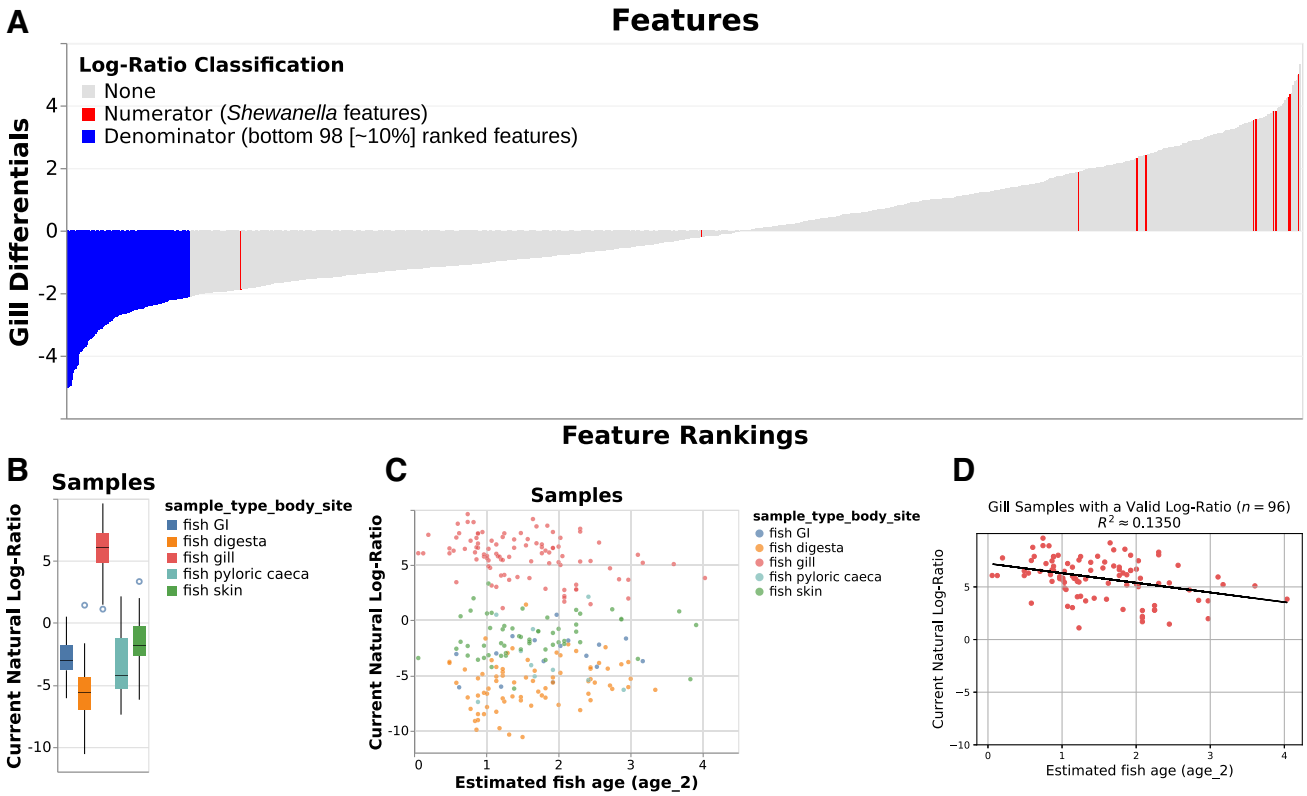

Figure 2. Various outputs from the case study (analogous to those in Figure 1) showing the log-ratio of classified *Shewanella* features to the bottom 98 ranked features for the gill differentials. (**A**) 'Rank plot' analogous to that shown in Figure 1A, with the selected numerator features (those classified as *Shewanella* spp.) colored in red and the selected denominator features (the bottom 98 ranked features for the gill differentials) colored in blue. (**B**) 'Sample plot' in boxplot mode, showing the selected log-ratios of samples by body site. A total of 252 samples are represented in this plot; as in Figure 1B, other samples were either filtered out upstream in the analysis or contained zeroes on at least one side of their log-ratio. (**C**) 'Sample plot,' showing a scatterplot of samples' selected log-ratios versus estimated fish age. Individual samples are colored by body site. As in panel B, only 252 samples are present. (**D**) Ordinary-least-squares linear regression ($R^2 \approx 0.1350$) between estimated fish age and the selected log-ratio for just the 96 gill samples shown in B and C, computed outside of Qurro as specified for Figure 1D.

(2); in particular, one question we might be interested in asking at this point is if *Shewanella* spp. are similarly abundant across other fish body sites. The remainder of this case study discusses a simple exploratory investigation in pursuit of an answer to this question, as well as to a few other questions that came up along the way.

Choosing a suitable 'reference frame'. The compositional nature of marker gene sequencing data means that we cannot simply compare the abundances of *Shewanella* across samples in this dataset alone; however, we can instead compare the log-ratio of *Shewanella* and other features in this dataset across samples (2).

For demonstrative purposes, we chose the taxonomic order *Synechococcales* as the denominator ('reference frame') for the first log-ratio shown here. Features in this dataset belonging to this order included sOTUs classified in the genera *Cyanobium*, *Prochlorococcus* and *Synechococcus*. These are common genera of planktonic picocyanobacteria found ubiquitously in marine surface waters (26). The expected stability of this group of features across samples in this dataset supports its use as a denominator here (2). Furthermore, as shown in Qurro's rank plot in Figure 1A, many *Synechococcales* features are relatively lowly ranked for the gill differentials; this gives additional reason to expect a comparative difference among gill samples for the *Shewanella*-to-*Synechococcales* log-ratio.

Qurro's computation of log-ratios. Currently, Qurro computes log-ratios between between arbitrary groups of N selected numerator features and D selected denominator features by, for each sample S, computing the log-ratio of the sums of the raw abundances of the numerator and denominator features:

$$\text{LogRatio}(S, N, D) = \ln\left(\sum_{n \in N} S_n\right) - \ln\left(\sum_{d \in D} S_d\right)$$

Computing log-ratios by summing feature abundances in this way, as opposed to taking the geometric mean of these abundances (e.g. as described in (27)) has benefits and downsides alike, as discussed in a recent preprint (28). One benefit is that this approach is relatively robust to highly sparse datasets like those commonly encountered in microbiome studies, since the presence of a zero-abundance feature in a group on one side of a sample's log-ratio does

not necessarily force this sample to have an invalid log-ratio. There are likely better alternatives to amalgamating feature abundances in this way, but this approach is useful for exploratory analysis nonetheless (and it is modifiable in Qurro's source code, should another method of amalgamation be desired in the future).

Relating log-ratios to sample metadata. Upon selecting a numerator and a denominator for a log-ratio (in this case, by searching through taxonomic annotations), Qurro updates the sample plot so that all samples' y-axis ('Current Natural Log-Ratio') values are equal to the value of the selected log-ratio for that sample. The *x*-axis field, color field, and scale types of these fields—along with other options—can be adjusted by the user interactively to examine the selected log-ratio from new perspectives.

Once the log-ratio of *Shewanella*-to-*Synechococcales* was selected, Figure 1B was produced by setting the sample plot x-axis to the categorical `sample_type_body_site` field and checking the 'Use boxplots for categorical data?' checkbox. The resulting boxplot shows that the *Shewanella*-to-*Synechococcales* log-ratio is relatively high in gill samples, compared with other body sites' samples (Figure 1B). This observation corroborates the initial study of this dataset on the frequency of *Shewanella* particular to the fish gill microbiome (14).

Qurro can visualize quantitative sample metadata, as well. Using this functionality, we can add additional perspectives to our previously-reached observation. Age has been discussed as a factor impacting the microbiota of fish gills in this and other datasets (14,25), and we use it here as an illustrative example of visualizing a quantitative metadata field alongside a log-ratio. By setting the *x*-axis field to the `age_2` metadata field (estimated fish age), changing the x-axis field scale type to 'Quantitative,' and setting the color field to `sample_type_body_site`, we get Figure 1C—a scatterplot showing the selected log-ratio viewed across samples by the estimated age of their host fish.

One trend that stood out to us in this scatterplot, and one of the reasons we chose age for this example, is that this plot contains an apparent negative correlation between the selected log-ratio and estimated fish age for gill samples. To support further investigation of patterns like this, Qurro can export the data backing the sample plot to a standard tab-separated file format—this file can then be loaded and analyzed in essentially any modern statistics software or programming language. This functionality was used to generate Figure 1D, in which we quantify and visualize this correlation for gill samples using ordinary-least-squares linear regression ($R^2 \approx 0.1008$). Although obviously not evidence of a causal relationship, this result opens the door for further investigation of this trend. One of many possible explanations for this observed trend is that the gills of younger fish are differentially colonized by *Shewanella* spp. and/or by *Synechococcales*; this may, in turn, be reflective of factors like vertical habitat use, immune development, or food choice.

Interrogating the 'multiverse' of reference frames. Prior literature has shown the impact that choices in data processing can have on a study's results, and on the corresponding 'multiverse' of datasets generated during this process (29). We submit that the choice of reference frame (denominator) in log-ratio analyses introduces a similar 'multiverse': for a set of n features, there are $O(2^n)$ possible subsets (2), so manually checking all possible reference frames for a given numerator is an intractable effort for the vast majority of datasets (although various heuristic methods have been proposed to address this sort of problem, e.g. (27)). In spite of this, the interactive nature of Qurro simplifies the task of validating results across reference frames.

Revisiting our analysis of *Shewanella* spp. in the gills of *S. japonicus*, there are multiple reasonable choices for reference frames. We chose *Synechococcales* mostly due to its expected ubiquity and stability across the marine samples in this dataset, but many other plausible choices exist.

In Figure 2, *we repeat the exact same analysis as in* Figure 1: but instead of using *Synechococcales* as the denominator of our log-ratio, we instead select the bottom $\sim 10\%$ (98/985) of features as ranked by gill differentials as the denominator (Figure 2A; Supplementary Data, section 2). Refreshingly, this log-ratio also shows clear 'separation' of gill samples from other body sites' samples in the dataset (Figure 2B), as well as a similar negative correlation between estimated fish age and this log-ratio for gill samples (Figure 2C and D) ($R^2 \approx 0.1350$). This serves as further evidence for our previous claims: although we still can't say for sure, we can now more confidently state that *Shewanella* spp. seem to be dominant in the gills of *S. japonicus*, and that *Shewanella* abundance in these fishes' gills seems to be negatively correlated with (estimated) fish age—since the trends shown in Figures 1B–D and 2B–D have held up across multiple log-ratios with *Shewanella* as the numerator.

Handling 'invalid' samples. It is worth noting that many samples—including all of the seawater samples in the Qurro visualization (Supplementary Data, section 3)–are not present in Figures 1B–D or 2B–D. If a given sample in Qurro cannot be displayed for some reason—for example, the sample has a zero in the numerator and/or denominator of the currently selected log-ratio—Qurro will drop that particular sample from the sample plot. Furthermore, to make sure the user understands the situation, Qurro will update a text display below the plot that includes the number and percentage of samples excluded for each 'reason.' This behavior helps users avoid spurious results caused by visualizing only a small proportion of a dataset's samples.

Using Qurro in practice

Since Qurro visualizations are essentially just web pages it is trivial to host them online, thus making them viewable by anyone using a compatible web browser. As an example of this we have made Qurro visualizations of various datasets, including the case study's, publicly available at https://biocore.github.io/qurro. We encourage users of Qurro to share their visualizations in this way, whenever possible, in order to encourage reproducibility and facilitate public validation of the conclusions drawn. Furthermore, we encourage readers of this paper to reconstruct Figures 1 and 2 and verify that this paper's claims are accurate.

CONCLUSION

Qurro serves as a natural 'first step' for users of modern differential abundance tools to consult in order to analyze feature rankings, simplifying the work needed to go from hypothesis to testable result. We have already found it useful in a variety of contexts, and it is our hope that others find similar value.

As more techniques for differentially ranking features become available, we believe that Qurro will fit in as a useful piece within the puzzles represented by modern 'omic studies.

DATA AVAILABILITY

All data used was obtained from study ID 11721 on Qiita. Deblur output artifact ID 56427 was used, in particular. Sequencing data is also available at the ENA (study accession PRJEB27458). Various Jupyter Notebooks and files used in the creation of this paper are available at https://github.com/knightlab-analyses/qurro-mackerel-analysis.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

The authors thank the anonymous reviewers for their thoughtful feedback and suggestions on the manuscript. The authors thank Julia Gauglitz, Shi Huang, Franck Lejzerowicz, Robert Mills, Justin Shaffer, Seth Steichen, Bryn Taylor and Yoshiki Vázquez-Baeza for helpful comments on the tool's functionality and design. The authors thank Gail Ackermann for assistance with study metadata. The authors thank Jerry Kennedy, Yoshiki Vázquez-Baeza, and Austin Swafford for assistance with funding information. The authors thank Jake VanderPlas, Dominik Moritz and the rest of the Altair/Vega development teams for answering questions regarding their software. Finally, the authors thank Sarah Allard, Tomasz Kościółek, Franck Lejzerowicz, Anupriya Tripathi and Yoshiki Vázquez-Baeza for help naming the tool.

FUNDING

This work is partly supported by IBM Research AI through the AI Horizons Network and the UC San Diego Center for Microbiome Innovation (to R.K. and M.W.F.); Joint University Microelectronics Program (JUMP)'s Center for Research on Intelligent Storage and Processing-in-memory (CRISP) [GI18518 to R.K. and M.W.F.]; University of California San Diego Computer Science and Engineering Department (to M.W.F.); University of California San Diego Frontiers of Innovation Scholars Program (to C.M.); National Science Foundation [GRFP DGE-1144086 to J.T.M.]; National Institute of Dental and Craniofacial Research through F31 Fellowship [1F31DE028478 to C.A.M.]; University of California San Diego Center for Microbiome Innovation through Microbial Sciences Graduate Research Fellowship (to J.J.M.).

Conflict of interest statement. None declared.

REFERENCES

1. Gloor,G.B., Macklaim,J.M., Pawlowsky-Glahn,V. and Egozcue,J.J. (2017) Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.*, **8**, 2224.
2. Morton,J.T., Marotz,C., Washburne,A., Silverman,J., Zaramela,L.S., Edlund,A., Zengler,K. and Knight,R. (2019) Establishing microbial composition measurement standards with reference frames. *Nat. Commun.*, **10**, 2719.
3. Fernandes,A.D., Reid,J.N., Macklaim,J.M., McMurrough,T.A., Edgell,D.R. and Gloor,G.B. (2014) Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome.*, **2**, 15.
4. Martino,C., Morton,J.T., Marotz,C.A., Thompson,L.R., Tripathi,A., Knight,R. and Zengler,K. (2019) A Novel Sparse Compositional Technique Reveals Microbial Perturbations. *mSystems*, **4**, e00016-19.
5. Aitchison,J. and Greenacre,M. (2002) Biplots of compositional data. *J. R. Stat. Soc. C.-Appl.*, **51**, 375–392.
6. Becker,R.A. and Cleveland,W.S. (1987) Brushing Scatterplots. *Technometrics.*, **29**, 127–142.
7. Bolyen,E., Rideout,J.R., Dillon,M.R., Bokulich,N.A., Abnet,C.C., Al-Ghalith,G.A., Alexander,H., Alm,E.J., Arumugam,M., Asnicar,F. *et al.* (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.*, **37**, 852–857.
8. Satyanarayan,A., Moritz,D., Wongsuphasawat,K. and Heer,J. (2016) Vega-Lite: A Grammar of Interactive Graphics. *IEEE. T. Vis. Comput. Gr.*, **23**, 341–350.
9. VanderPlas,J., Granger,B., Heer,J., Moritz,D., Wongsuphasawat,K., Satyanarayan,A., Lees,E., Timofeev,I., Welsh,B. and Sievert,S. (2018) Altair: Interactive Statistical Visualizations for Python. *J. Open Source Softw.*, **3**, 1057.
10. McDonald,D., Clemente,J.C., Kuczynski,J., Rideout,J.R., Stombaugh,J., Wendel,D., Wilke,A., Huse,S., Hufnagle,J., Meyer,F. *et al.* (2012) The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience.*, **1**, 7.
11. Van Der Walt,S., Colbert,S.C. and Varoquaux,G. (2011) The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.*, **13**, 22–30.
12. McKinney,W. (2010) Data Structures for Statistical Computing in Python. In: Van,Der Walt S and Millman,J (eds). *Proceedings of the 9th Python in Science Conference*, SciPy, Austin, pp. 51–56.
13. Satyanarayan,A., Russell,R., Hoffswell,J. and Heer,J. (2015) Reactive Vega: A Streaming Dataflow Architecture for Declarative Interactive Visualization. *IEEE. T. Vis. Comput. Gr.*, **22**, 659–668.
14. Minich,J., Petrus,S., Michael,J.D., Michael,T.P., Knight,R. and Allen,E. (2019) Temporal, environmental, and biological drivers of the mucosal microbiome in a wild marine fish, *Scomber japonicus*. bioRxiv doi: http://dx.doi.org/10.1101/721555, 01 August 2019, preprint: not peer reviewed.
15. Kluyver,T., Ragan-Kelley,B., Pérez,F., Granger,B.E., Bussonnier,M., Frederic,J., Kelley,K., Hamrick,J.B., Grout,J., Corlay,S. *et al.* (2016) Jupyter Notebooks-a publishing format for reproducible computational workflows. In: Loizides,F and Schmidt,B (eds). *Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing*. IOS Press, Amsterdam, pp. 87–90.
16. Minich,J.J., Zhu,Q., Janssen,S., Hendrickson,R., Amir,A., Vetter,R., Hyde,J., Doty,M.M., Stillwell,K., Benardini,J. *et al.* (2018) KatharoSeq enables high-throughput microbiome analysis from low-biomass samples. *mSystems*, **3**, e00218-17.
17. Caporaso,J.G., Kuczynski,J., Stombaugh,J., Bittinger,K., Bushman,F.D., Costello,E.K., Fierer,N., Pena,A.G., Goodrich,J.K., Gordon,J.I. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods.*, **7**, 335–336.
18. Amir,A., McDonald,D., Navas-Molina,J.A., Kopylova,E., Morton,J.T., Xu,Z.Z., Kightley,E.P., Thompson,L.R., Hyde,E.R., Gonzalez,A. *et al.* (2017) Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*, **2**, e00191-16.

19. Gonzalez,A., Navas-Molina,J.A., Kosciolek,T., McDonald,D., Vázquez-Baeza,Y., Ackermann,G., DeReus,J., Janssen,S., Swafford,A.D., Orchanian,S.B. *et al.* (2018) Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods.*, **15**, 796–798.
20. Bokulich,N.A., Kaehler,B.D., Rideout,J.R., Dillon,M., Bolyen,E., Knight,R., Huttley,G.A. and Caporaso,J.G. (2018) Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome.*, **6**, 90.
21. Quast,C., Pruesse,E., Yilmaz,P., Gerken,J., Schweer,T., Yarza,P., Peplies,J. and Glöckner,F.O. (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.
22. Parada,A.E., Needham,D.M. and Fuhrman,J.A. (2016) Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ. Microbiol.*, **18**, 1403–1414.
23. Apprill,A., McNally,S., Parsons,R. and Weber,L. (2015) Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquat. Microb. Ecol.*, **75**, 129–137.
24. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
25. Pratte,Z.A., Besson,M., Hollman,R.D. and Stewart,F.J. (2018) The gills of reef fish support a distinct microbiome influenced by host-specific factors. *Appl. Environ. Microb.*, **84**, e00063-18.
26. Scanlan,D.J., Ostrowski,M., Mazard,S., Dufresne,A., Garczarek,L., Hess,W.R., Post,A.F., Hagemann,M., Paulsen,I. and Partensky,F. (2009) Ecological Genomics of Marine Picocyanobacteria. *Microbiol. Mol. Biol. Rev.*, **73**, 249–299.
27. Rivera-Pinto,J., Egozcue,J., Pawlowsky-Glahn,V., Paredes,R., Noguera-Julian,M. and Calle,M. (2018) Balances: a New Perspective for Microbiome Analysis. *mSystems*, **3**, e00053-18.
28. Quinn,T. and Erb,I. (2020) Amalgams: data-driven amalgamation for the reference-free dimensionality reduction of zero-laden compositional data. bioRxiv doi: https://doi.org/10.1101/2020.02.27.968677, 28 February 2020, preprint: not peer reviewed.
29. Steegen,S., Tuerlinckx,F., Gelman,A. and Vanpaemel,W. (2016) Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.*, **11**, 702–712.
30. Hunter,J.D. (2007) Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.